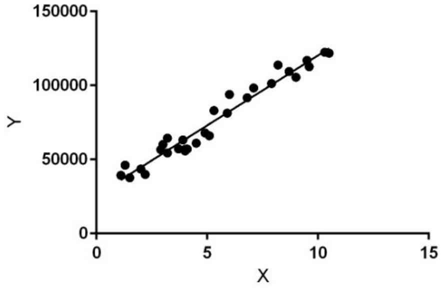
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
    - Based on the analysis of categorical variables from the dataset it can be said that these variables have a significant impact on the dependent / target variable (cnt).
    - To list some examples:
        - Considering the correlation matrix we can see that year and count have a moderate positive correlation indicating that there has been an increase in bike rentals from the year 2018 to 2019. Note: year (0: 2018, 1:2019)
        - Considering the p-values and VIF values of variables like "season" (Spring, Summer, Winter), "weathersit" (Snowy, Sunny) etc., they appear to have significant impact on the variations in the target variable.
        - Finally, inclusion of all these categorical variables gave us a good R-squared value of 0.805 which justifies the significance of these variables.
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
    - When we create dummy variables from categorical variables we use 'drop_first=True' to leave out one category (known as reference category) because not doing so will make the first column a redundant feature and introducing multicollinearity. This makes the analysis more accurate and easy to understand.
    - In conclusion, if you have k categorical variables, you can have k-1 dummy variables.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
    - On first look, 'registered' is the variable with the highest correlation with 'cnt'. But after data analysis when we drop 'registered' as it is a variable that won't be considered for model we can see that **'temp'** has the highest correlation with target variable 'cnt'.
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    - **Linear relationships**: Identifying linear relationships between the independent variable and the target variable. This was done using Scatter plots.
    - **Multicollinearity**: Calculating the VIF (Variance Inflation Factor) to check for multicollinearity among independent variables. High VIF value (>5) indicates high multicollinearity which affects the model's accuracy and performance.
    - Validation of **p-value**
    - **Residual Analysis**: Checking for residuals which is the difference between actual and predicted values. The residuals should be scattered around 0.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- **'yr' (Year)** => 0: 2018, 1:2019. It has a coefficient of 0.2350, which indicates a positive relationship with the demand for bike rentals which seems to have increased from 2018 to 2019.
- **temp (Temperature):** This feature has a coefficient of 0.4673, indicating a positive relationship with the demand for bike rentals. As we also saw in the correlation matrix 'temp' had the highest correlation with our target variable 'cnt'.
- **Sunny**: This was a categorical variable from '**weathersit'** having value 1 which represents **'Clear, Few clouds, Partly cloudy, Partly cloudy'** which was mapped to 'Sunny'. It has a coefficient of 0.0789 and indicates a positive relationship with the demand for bike rentals.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
    - We have two categories of Machine learning models: Supervised learning and Unsupervised learning. Linear regression is a machine learning model based on supervised learning.
    - Supervised learning is a method which allows us to use historical / past data with labels for model training. This past data is divided into 2 parts in supervised learning method i.e., 1) Training data which is used for training or enabling the model to learn and 2) Testing data which is used by the trained model for predictions and evaluations of the model.
    - Linear regression is an algorithm which is used for predicting a numerical value (continuous values) based on the relationship between a dependent (target) variable and one or more independent variables. There are 2 types of Linear regression algorithms
        - Simple linear regression: When number of independent variables is 1

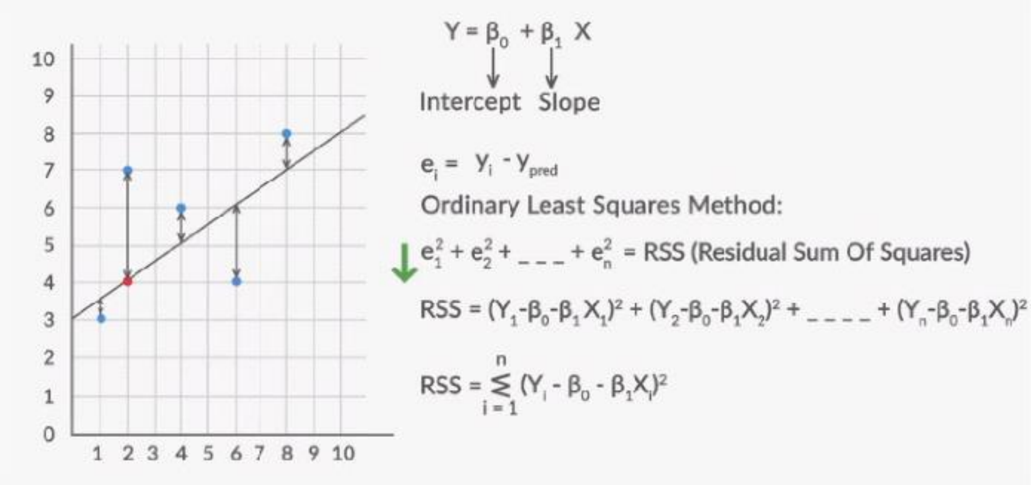


Linear Regression

        - Multiple linear regression: When number of independent variables is 2 or more.
    - The goal of this algorithm is to find the best linear equation which can predict the value of target / dependent variable based on independent variables. This equation

provides a straight line that represents relationship between the dependent and independent variables.

  - Assumptions of Simple Linear regression: Linear regression has some conditions or requirements that should be met for the model to provide accurate and reliable results.
    - Linear relationship between X and y.
    - Normal distribution of error terms.
    - Independence of error terms.
    - Constant variance of error terms.
  - The equation of the best fit regression line Y = $\beta_0 + \beta_1 X$ can be found by minimising the cost function (RSS – Residual Sum squares) which is equal to sum of squares of residual for each data point in the plot.


10
9
8
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10
$Y = \beta_0 + \beta_1 X$
Intercept
Slope
$e_i = y_i - y_{pred}$
Ordinary Least Squares Method:
$e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)
$RSS = (Y_1-\beta_0-\beta_1X_1)^2 + (Y_2-\beta_0-\beta_1X_2)^2 + \_\_\_\_ + (Y_n-\beta_0-\beta_1X_n)^2$
$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$

    -
    - This can be done using 2 methods: Differentiation and Gradient descent method
  - The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$
    - RSS: Residual Sum of Squares
    - TSS: Total Sum of Squares
- Assumptions of Multiple Linear Regression:
    - Overfitting - When the model becomes too complex and gives good results in training data but fails when working with testing data.
    - Multicollinearity - Identifying if there is any dependency within the independent variables and remove them in order to avoid redundancy.
    - Feature selection –Picking the most important and relevant features for the model and dropping the redundant features.
- R-squared and Adjusted R-squared are used to evaluate how well the regression model fits the data. Both of these help in assessing the goodness of fit of the model. While R-squared shows the overall proportion of variance explained, Adjusted R-squared helps to prevent overfitting by considering the complexity of the model. In

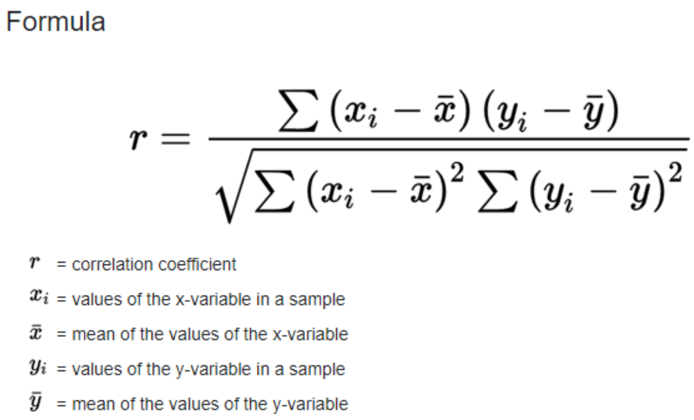
general, higher values of both these metrics indicate a better model. Along with these metrics we should also consider other influencing factors such as Co-efficients, p-value, VIF, domain knowledge etc.,

2. Explain the Anscombe’s quartet in detail. (3 marks)

- Anscombe’s quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph.
- It demonstrates the importance of visualizing data and shows that considering summary statistics alone can be misleading.
- The four datasets that make up Anscombe’s quartet each include 11 x-y pairs of data. When we plot these we can see that each dataset seems to have a unique connection between x and y, with unique variability patterns. Despite these variations, each dataset has the same summary statistics, meaning same x and y mean and variance, correlation coefficient and linear regression line.
- The takeaway from Anscombe's quartet is that summary statistics like means, variances and correlations can hide important details in data. Although they seem to have similar statistic properties they show a different relationship when visualized with graphs which highlights the need for data visualization to gain better insights on the data.

3. What is Pearson’s R? (3 marks)

- Pearson's R also known as "correlation coefficient" is a number that tells us how much two sets of numbers like x and y for example are related to each other. It gives a value between -1 and +1. If the correlation coefficient is close to +1, it means that the two sets of numbers have a strong positive relationship meaning when one goes up, the other tends to go up as well. If the correlation coefficient is close to -1 it means that they have a strong negative relationship i.e., when one goes up the other tends to go down. If the correlation coefficient is close to 0 it means that there isn't much relationship between the two sets of numbers. By doing so Pearson's R helps us understand how two sets of data are connected.



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

-
- You can also use numpy library’s corrcoef function to calculate this value in Python.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

- Scaling is an important step which brings all the feature values to the same scale (within a particular range).
- Some commonly used scaling techniques are:
    - Min Max Scaling (Normalized scaling)
    - Standardization
        - (mean=0 and sigma=1) => (x-mu)/sigma
- Importance of scaling:
    - Variables containing large values can dominate over the small values and model might end up giving more importance to features with these large values. Also, the categorical variables which are encoded to binary form (0,1) so not performing scaling on the dataset where data ranges are spread out will result in the model favoring variables with large values and giving inaccurate predictions.
    - This can improve the performance of the algorithms
- Normalized scaling vs Standardized scaling
    - Normalized scaling
        - This method transforms the data such that all values fall within the range of 0 to 1.
        - This approach also handles outliers in data.
        - Formula => (x- xmin)/(xmax-xmin)
        - This approach is useful when we wish to maintain relative relationships between variables.
    - Standardized Scaling
        - This method transforms data such that it has a mean of 0 and standard deviation of 1.
        - Formula => (x – mean(x)) / std(x)
        - This approach is useful when we want to make sure all features have the same mean and standard deviation.
        - Helpful in algorithms that use gradient descend.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

- VIF (Variance Inflation Factor) is basically a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated/dependent on each other, which can cause issues in interpreting the effects of individual variables.

- When the VIF value is infinite it basically means that there's a perfect linear relationship between variables. This happens when one or more variables in the model can be exactly predicted using a combination of other variables.
- This infinite VIF value indicates an extreme multicollinearity and it usually occurs due to data errors, mathematical errors or scenarios where variables are linearly dependent on each other. In this case we need to revisit the model and remove correlated/dependent variables and restructure the model for better stability and accuracy.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

- A Q-Q plot also known as the Quantile-Quantile plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution like the normal distribution. It helps us compare the quantiles of us dataset's values with the quantiles of a theoretical distribution. In simple terms, it shows if the data behaves like you expect it to.
- Use and Importance of Q-Q Plot in Linear Regression:
    - Checking for the Normality Assumption: In linear regression, one of the assumptions is that the residuals (the differences between observed and predicted values) should follow a normal distribution. If this assumption is violated, it can affect the performance of the regression model. This is where a Q-Q plot comes into picture to check whether the residuals are normally distributed.
    - Outliers detection: Q-Q plots can identify outliers that might not be very obvious in other types of graphs or plots. If the data points deviate significantly from the expected values in the Q-Q plot, it indicates the presence of outliers in the data.
    - Model validation: Linear regression assumes that the errors or residuals are normally distributed. A Q-Q plot helps you assess if these assumptions hold true. If points on the Q-Q plot deviate from the straight line, it means that the assumptions might not be met.